Collaborative Human-AI (CHAI): Evidence-Based Interpretable Melanoma Classification in Dermoscopic Images

Noel C. F. Codella[1], Chung-Ching Lin[1], Allan Halpern[2], Michael Hind[1], Rogerio Feris[1], and John R. Smith[1]

[1]IBM T.J. Watson Research Center, Yorktown Heights, NY, 10598, USA
[2]Memorial Sloan-Kettering Cancer Center, New York, NY, 10065
[1]{nccodell,cclin,hindm,rferis,jsmith}@us.ibm.com
[2]{halperna}@mskcc.org

Abstract. Automated dermoscopic image analysis has witnessed rapid growth in diagnostic performance. Yet adoption faces resistance, in part, because no evidence is provided to support decisions. In this work, an approach for evidence-based classification is presented. A feature embedding is learned with CNNs, triplet-loss, and global average pooling, and used to classify via kNN search. Evidence is provided as both the discovered neighbors, as well as localized image regions most relevant to measuring distance between query and neighbors. To ensure that results are relevant in terms of both label accuracy and human visual similarity for any skill level, a novel hierarchical triplet logic is implemented to jointly learn an embedding according to disease labels and non-expert similarity. Results are improved over baselines trained on disease labels alone, as well as standard multiclass loss. Quantitative relevance of results, according to non-expert similarity, as well as localized image regions, are also significantly improved.

Keywords: deep learning, evidence, explainable, interpretable, triplet-loss, global average pooling, weighted activation maps, dermoscopy, melanoma

1 Introduction

In the past decade, advancement in computer vision techniques has been facilitated by both large-scale datasets and deep learning approaches. Now this trend is influencing dermoscopic image analysis, where the International Skin Imaging Collaboration (ISIC) has organized a large public repository of high quality annotated images, referred to as the ISIC Archive (http://isic-archive.com). From this repository, snapshots of the dataset have been used to host two consecutive years of benchmark challenges [1, 2], which have increased interest in the computer vision community [2–6], and supported the development of methods that surpassed the diagnostic performance of expert clinicians [2–4]. However, despite these advancements, deployment to clinical practice remains problematic, in part, because most systems lack evidence for predictions that can be interpreted by users of varying skill.

Recent works have attempted to provide various forms of evidence for decisions. Methods to visualize feature maps in neural networks were introduced in 2015 [7], facilitating better understanding of the behavior of networks, but not justifying predictions made on specific image inputs. Global average pooling approaches have been proposed [8], which get closer to justifying decisions on specific image inputs by indicating importance of image regions to those decisions, but fail to provide specific evidence behind the classifications.

An extensive body of prior work around content-based image retrieval (CBIR) is perhaps the most relevant toward providing classification decisions with evidence [9–13]. Early approaches relied on low-level features and bag-of-visual words, [9–11], but suffered from the "semantic gap": feature similarity did not necessarily correlate to label similarity. Later approaches have used deep neural networks to learn an embedding for search, reducing semanic gap issues [13]. However, such methods have still suffered from a "user-gap": what an embedding learns to consider as similar from disease point-of-view does not necessarily correlate with human measures of similarity. In addition, users cannot determine what spatial regions of images contributed most to distance measures.

Specific to the domain of dermoscopic image analysis, one work proposed to learn and localize clinically discriminative patterns in images [5]; however, this output can only be verified by experts who know how to identify the patterns. In addition, classifier decision localization has been proposed for multimodal systems [14]; however, localization information alone isn't sufficient as evidence for classification decisions.

In this work, a solution for a Collaborative Human-AI (CHAI) dermoscopic image analysis system is presented. In order to facilitate interpretability of evidence by clinical staff of any skill level, this approach 1) introduces a novel hierarchical triplet loss to learn an embedding for k-nearest neighbor search, optimized jointly from disease labels as well as non-expert human similarity, and 2) provides localization information in the form of *query-result activation map pairs*, which designate regions in query and result images used to measure distance between the two. Experiments demonstrate that the proposed approach improves classification performance in comparison to models trained on disease labels alone, as well as models trained with classification loss. The relevancy of results, according to non-expert similarity, are also significantly improved.

2 Methods

2.1 Triplet-Loss with Global Average Pooling

The proposed embedding framework is displayed in Fig. 1a. A triplet loss structure [15] is combined with penultimate global average pooling layers [8] to learn a discriminative feature embedding that supports activation localization. AlexNet, including up to the "conv5" layer, is used as the CNN.

In order to train, 3 deep neural networks with shared weights across 3 input images (x^a, x^b, x^c) produce feature embeddings $(f(x^a), f(x^b), f(x^c))$. The

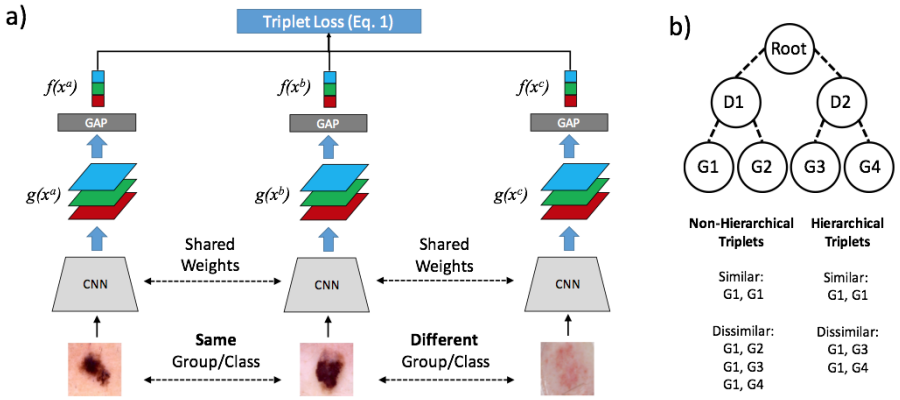

Fig. 1. a) Proposed triplet loss framework with global average pooling (GAP) architecture. b) Top: Visual example of proposed hierarchical annotation groups. The first level grouping is by disease label (D1-2), and the second level by human visual similarity (G1-4). Bottom: Example triplet logic is shown as pairing between groups.

following objective function over those embeddings provides the gradient for backpropagation:

$$L = max\left[0, l + D(f(x^a), f(x^b)) - \frac{1}{2}(D(f(x^a), f(x^c)) + D(f(x^b), f(x^c)))\right] \quad (1)$$

where $D()$ is a distance metric (squared Euclidean distance), l is a constant representing the margin (set to 1), x^a and x^b are considered similar inputs, and x^c is a dissimilar input.

The feature embedding is comprised of a global average pooling (GAP) layer to support generation of a *query-result activation map pair*, which highlights regions of pairs of images that contributed most toward the distance measure between them. This is done by combining the feature layer activation maps prior to global average pooling into a single grayscale image, weighted by the squared differences between two image feature embeddings:

$$A^q(i,j) = \sum_{z=0}^{d} g_z(x^q, i, j)) \cdot (f_z(x^q) - f_z(x^r))^2 \quad (2)$$

where $A^q(i,j)$ is the query activation map (QAM), $g_z(x,i,j)$ is the z^{th} filter bank before global average pooling, d is the dimensionality of the filter bank, x^q is the query image, x^r is a search result image, and:

$$f_z(x) = \frac{1}{n^2} \sum_{i=0}^{n} \sum_{j=0}^{n} g_z(x,i,j) \quad (3)$$

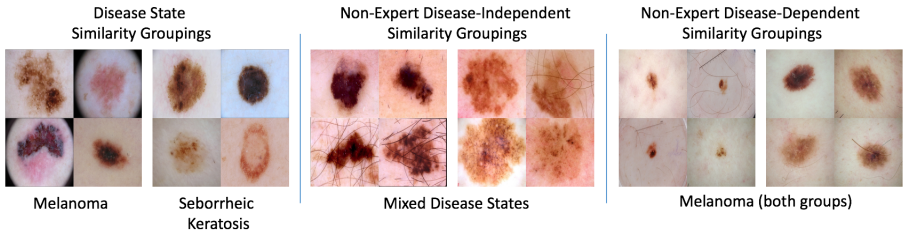

Fig. 2. Example groups by disease category from the ISIC database (left), by non-expert similarity disregarding disease diagnosis (center), and by non-expert similarity constrained within disease groups (right).

is the z^{th} feature embedding element. The result activation map (RAM) A^r in the query-result pair is likewise computed as in Eq. 2, where $g_z(x^q, i, j)$ is replaced with $g_z(x^r, i, j)$.

2.2 Hierarchical Triplet Selection Logic

An example of the hierarchical triplet selection logic is shown in Fig. 1b. Given visually similar groups annotated under disease labels, a hierarchical selection process pairs images as similar if they are siblings within the same group under a disease parent. Dissimilar images include images from other disease states, but exclude cousin images (images within the same disease, but different similarity group). A non-hierarchical selection process takes dissimilar images from any other group, including cousins.

2.3 Experimental Design

The 2017 International Skin Imaging Collaboration (ISIC) challenge on Skin Lesion Analysis Toward Melanoma Detection [1] dataset is used for experimentation. This is a public dataset consisting of 2000 training dermoscopic images and 600 test images. Experiments on this data compare between the following 6 feature embeddings for kNN classification:

Baseline: The first is the 4096 dimensional fc6 feature embedding layer of the AlexNet architecture trained on the CASIA-WebFace dataset, described in prior work [15]. This is used as the baseline as it is one of the only human-skin focused pre-trained networks currently available.

BaselineFT: Baseline 4096 is fine-tuned for disease labels using standard multiclass accuracy loss. This method represents one of the most common approaches for generating embeddings for KNN classification in practice.

Disease: This is a 1024 dimensional CHAI feature embedding, learned from disease labels on the training data partition of the ISIC dataset, fine-tuned from the baseline.

Joint: This is a CHAI feature embedding jointly fine-tuned from baseline using disease labels, as well as non-expert human similarity groupings, consisting

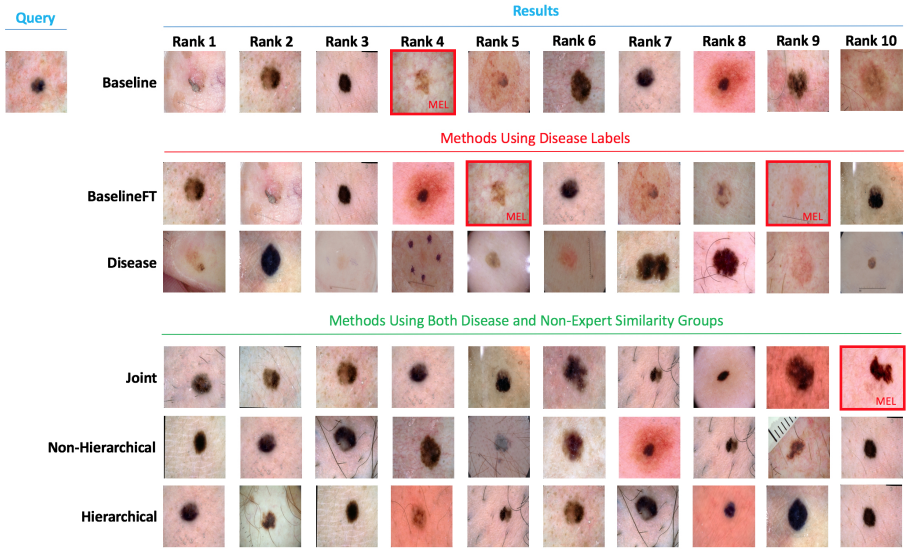

Fig. 3. Example search results across systems, displayed according to similarity rank, with rank 1 being the most similar image in the training dataset. Red borders signify instances of melanoma.

of 1700 images pulled from the ISIC Archive (excluding test images), annotated into 37 distinct groups. The annotator was not given disease labels, and thus may mix diseases within groups. Example groups are shown in Fig. 2.

Hierarchical: This is a CHAI feature embedding fine-tuned from the disease model using human similarity groups that are dependent on disease labels. All 2000 images and 600 test images were annotated from the 2017 ISIC challenge dataset, partitioned into 20 groups of similar images under melanoma, 12 groups under seborrheic keratosis, and 15 groups under benign nevus, according to a non-expert human user. Because this type of data is difficult to annotate, only 1000 training images were used for fine-tuning. The remainder of the data was used for evaluation. Examples of these groups are shown in Fig. 2. Triplets were selected based on hierarchical logic.

Non-Hierarchical: To isolate the effects of hierarchical logic, and disease labels being provided to the annotator, the hierarchical groups are used to create triplets using non-hierarchical logic: dissimilar images are selected from any other group, including cousins.

Most learning parameters are kept consistent with prior art [15], including the activation map feature dimensionality of 1024 [8], batch size 128, momentum of 0.9, "step" learning rate policy, learning rate for transferred weights (0.00001), and learning rate for randomly initialized GAP layer (0.01). For BaselineFT, a learning rate of 0.01 was used for fc8, 0.001 for fc7 and fc6, and 0.00001 for earlier layers. For all triplet experiments, 150,000 triplets were randomly generated for training, and 50,000 triplets for validation.

	Baseline	BaselineFT	Disease	Joint	Non-Hierarchical	Hierarchical
AUC k3	0.663	0.700	**0.734**	0.704	0.713	0.729
AUC k5	0.675	0.714	0.744	0.738	0.743	**0.756**
AUC k10	0.681	0.709	0.757	0.754	0.749	**0.774**
AUC k20	0.712	0.745	0.775	0.752	0.769	**0.783**
AUC k40	0.691	0.742	0.776	0.760	0.776	**0.786**
REL k3	0.942	1.005	0.865	1.048	**1.212**	1.125
REL k5	1.505	1.608	1.412	1.678	**1.958**	1.872
REL k10	2.875	3.027	2.632	3.147	**3.793**	3.658
REL k20	5.470	5.772	4.903	6.067	**7.300**	6.968
REL k40	10.283	10.703	9.125	11.507	**13.958**	13.333
JA	NA	NA	0.176	0.201	0.193	**0.208**

Table 1. Melanoma Classification AUC for each method and number of neighbors (k), followed by number of results matching human similarity relevancy (REL), and Jaccard (JA) of QAM against segmentation ground truth.

The area under receiver operating characteristic (ROC) curve (AUC) is used to measure melanoma classification performance on the dataset, according to average vote among returned nearest neighbors. The hierarchical similarity annotations were used to measure the average number of results matching non-expert human relevancy (REL) across all experiments. Finally, the quality of query activation maps are quantitatively measured by comparing the maps against ground truth segmentation according to Jaccard (JA).

3 Results

Table 1 shows the measured AUC for each model type and variable number of neighbors (k), the number of results matching non-expert human similarity relevance (REL), and the Jaccard of the query activation maps as judged against ground truth segmentations. For comparison, standard classification output from multi-class loss used to train *BaselineFT* produces an AUC of 0.772. The top AUC measured for the challenge was 0.874 [5].

For $k = 3$, *Disease* achieved the highest AUC. Surprisingly, at $k = 20, 40$, *Disease* outperforms the classification output of *BaselineFT* (0.772 AUC). For all other values of k, the *Hierarchical* triplet loss embedding achieved the highest performance. At $k = 40$, these performance numbers were comparable with predictive systems submitted to the challenge (rank 11 out of 23 submissions). The *Hierarchical* triplet loss also achieved the second highest number of human similarity relevant results. While the *Non-Hierarchical* method achieved the highest degree of human similarity relevant results, this came at the marginal cost of some classification performance in comparison to *Hierarchical* triplets. However, *Non-Hierarchical* has still matched the classification performance of *Disease*, and outperformed the standard multiclass loss of *BaselineFT*. *Joint* also showed improvements to relevance of human similarity in comparison to *Disease*, but

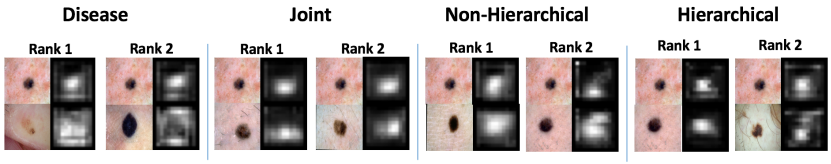

Fig. 4. Example query-result activation pairs for search results. In each group of 4 images: *Top-Left:* query image. *Top-Right:* query activation map. *Bottom-Left:* search result. *Bottom-Right:* search result activation map.

suffered a more harsh penalty to classification performance in comparison to *Hierarchical* and *Non-Hierarchical*.

Representative search results can be inspected in Fig. 3. One can observe here how *Disease*, trained directly on triplets from disease labels, does not translate into the most "relevant" results by human measure: clearly, rank 3 has returned a hypo-pigmented lesion for a pigmented lesion query. In contrast, *Joint*, while maintaining a robust improvement in AUC measures over *Baseline* and *BaselineFT*, has additionally learned to balance disease similarity with a more human measure of similarity. *Hierarchical* has both managed to improve classification performance and human similarity.

Finally, example query-result activation map pairs are shown in Fig. 4. Interestingly, *Disease* learned to examine a broad image extent during comparisons (even potentially irrelevant areas of images), whereas for the models trained with human measures of similarity, the systems have learned to focus more to the localized lesion area. This is confirmed in the over 10% quantitative improvement in Jaccard index comparing to ground truth lesion segmentations, as shown in Table 1.

4 Conclusion

In conclusion, "CHAI", a Collaborative Human-AI system to perform comprehensive evidence-based melanoma classification in dermoscopic images has been presented. Evidence is provided as both the nearest neighbors used for classification, as well as query-result activation map pairs that visualize regions of the images contributing most toward a distance computation. Using a novel hierarchical triplet loss, non-expert human similarity is used to tailor the feature embedding to more closely approximate human judgments of relevance, while simultaneously improving classification performance and the quality of the activation maps. Future work must be carried-out to determine 1) whether the method has the potential to improve adoption, 2) how to improve classification performance to better compete with other black-box systems, and 3) whether passive user interaction with a deployed system can be used for training (for example, from a user clicking on a specific evidence result) to improve classification performance and relevance over time with continued use.

References

1. Codella N, et al. Skin Lesion Analysis toward Melanoma Detection: A Challenge at the International Symposium on Biomedical Imaging (ISBI) 2017, hosted by the International Skin Imaging Collaboration (ISIC). IEEE International Symposium of Biomedical Imaging (ISBI) 2018.
2. Marchetti M, et al. "Results of the 2016 International Skin Imaging Collaboration International Symposium on Biomedical Imaging challenge: Comparison of the accuracy of computer algorithms to dermatologists for the diagnosis of melanoma from dermoscopic images". J Am Acad Dermatol. 2018 Feb;78(2):270-277
3. Codella NCF, Nguyen B, Pankanti S, Gutman D, Helba B, Halpern A, Smith JR. "Deep learning ensembles for melanoma recognition in dermoscopy images" In: IBM Journal of Research and Development, vol. 61, no. 4/5, 2017.
4. Esteva A,Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S. "Dermatologist-level classification of skin cancer with deep neural networks". Nature, vol 542, pp 115118. 2017.
5. Menegola A, Tavares J, Fornaciali M, Li LT, Avila S, Valle E. "RECOD Titans at ISIC Challenge 2017". 2017 International Symposium on Biomedical Imaging (ISBI) Challenge on Skin Lesion Analysis Towards Melanoma Detection. Available: https://arxiv.org/pdf/1703.04819.pdf
6. Diaz, I.G. Incorporating the Knowledge of Dermatologists to Convolutional Neural Networks for the Diagnosis of Skin Lesions. 2017 International Symposium on Biomedical Imaging (ISBI) Challenge on Skin Lesion Analysis Towards Melanoma Detection. Available: https://arxiv.org/abs/1703.01976
7. Yosinki, J., Clune, J., Nguyen, A., Fuchs, T., Lipson, H. Understanding Neural Networks Through Deep Visualization. In: Deep Learning Workshop of International Conference on Machine Learning (ICML) 2015.
8. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A. Learning Deep Features for Discriminative Localization. In: Computer Vision and Pattern Recognition (CVPR) 2016.
9. Akgul, C.B., Rubin, D.L., Napel, S., Beaulieu, C.F., Greenspan, H., Acar, B. Content-Based Image Retrieval in Radiology: Current Status and Future Directions. Journal of Digital Imaging, Vol 24, No 2 (April), 2011: pp 208Y222
10. Muller, H., KalpathyCramer, J., Caputo, B., Syeda-Mahmood, T., Wang, F. Overview of the First Workshop on Medical ContentBased Retrieval for Clinical Decision Support at MICCAI 2009.
11. Ballerini, L., Fisher, R., Rees, J.: A query–by–example content–based image retrieval system of non-melanoma skin lesions. In Medical Content-based Retrieval for Clinical Decision Support (MCBRCDS), LNCS, vol. 5853, 2009.
12. Li, Z., Zhang, X., Muller H, Zhang, S. Large-scale retrieval for medical image analytics: A comprehensive review. Medical Image Analysis 43 (2018) 6684
13. Chung Y.A., Weng, W.H. "Learning Deep Representations of Medical Images using Siamese CNNs with Application to Content-Based Image Retrieval. NIPS 2017 Workshop on Machine Learning for Health (ML4H)
14. Ge, Z., Demyanov S., Chakravorty, R., Bowling, A., Garnavi, R. Skin Disease Recognition Using Deep Saliency Features and Multimodal Learning of Dermoscopy and Clinical Images. MICCAI 2017.
15. Zhang, S., Gong, Y., Huang, J.B., Lim, J., Wang, J., Ahuja, N., Yang, M.H. Tracking Persons-of-Interest via Adaptive Discriminative Features. In: European Conference on Computer Vision (ECCV) 2016.